

Why database searches

- Gene finding
- Assigning likely function to a gene.
- Identifying regulatory elements
- Understanding genome evolution.
- Assisting in sequence assembly
- Finding relations between genes

Search engines:

3 main components:

- Scoring function
- Algorithm
- Statistical model to recover significant results

Important issue: speed

Local alignment

- Local alignment seeks similar segments of unspecified length from the 2 sequences being compared.
- Rigorous method is local dynamic programming (last class), time is proportional to the product of lengths of sequences it compares.
- BLAST is linear time *heuristic* algorithm.

BLAST

- Basic Local Alignment Search Tool – a family of most popular sequence search program including: Basic BLAST, Gapped BLAST, Psi - BLAST
- **Main idea (basic BLAST):** Homologous sequences are likely to contain a short high scoring similarity region **a hit**. Each hit gives a seed that BLAST tries to extend on both sides

Some BLAST terminology

word – substring of a sequence

word pair – pair of words of the same length.

score of a word pair – score of the gapless alignment of the two words:

V A L M R

V A K N S Score = $-4 + 3 + -4 + -3 + -1 = -9$

(PAM120)

HSP – high scoring sequence pair.

Main steps of BLAST

- **Parameters:** **w** = length of a hit; **T** = min. score of a hit (for proteins: w=3, T=13 (BLOSUM62))
- **Step 1:** Given query sequence Q, compile the list of possible words which form with words in Q high scoring word pairs.
- **Step 2:** Scan database for exact matching with the list of words compiled in step 1.
- **Step 3:** Extending hits from step 2.
- **Step 4:** Evaluating significance of extended hits from step 3.

Step 1: Find high scoring words

- For every word x of length w in Q make a list of words that when aligned to x score at least T .
- Example: Let $x=AIV$ then score for AIA is $5+5+0$ (dropped) and for AII $5+5+4$ (taken)
- Number of words in the list depends on w and T , and is much less than 20^3 (typically about 50)

Step 1

MVRERKCILCHIVY**GSK**KEMDEHMRSMMLHHRELENLKGRDIS

Query word, W=3 for proteins ↓

(W=11 for nucleotides)

Word Score (BL-62)

GSK 15

GAK 12

GNK 12

GTK 12

GSR 12

GDK 11

GQK 11

GEK 11

GGK 11

GKK 11

GSQ 11

GSE 11

Step 2 – Finding hits

- Scan database for exact matching with the list of words compiled in step1 :
- This can be done efficiently using techniques as hash table (requires preprocessing of a data base)

Step 2

MVRERKCILCHIVY**GSK**KEMDEHMRSMMLHHRELENLKGRDIS

Query word, W=3



Word Score (BL-62)

GSK	15	GAK	12	GNK	12
		GTK	12	GSR	12
GDK	11	GQK			
11				GEK	11
GGK	11			GKK	11
GSQ	11			GSE	11

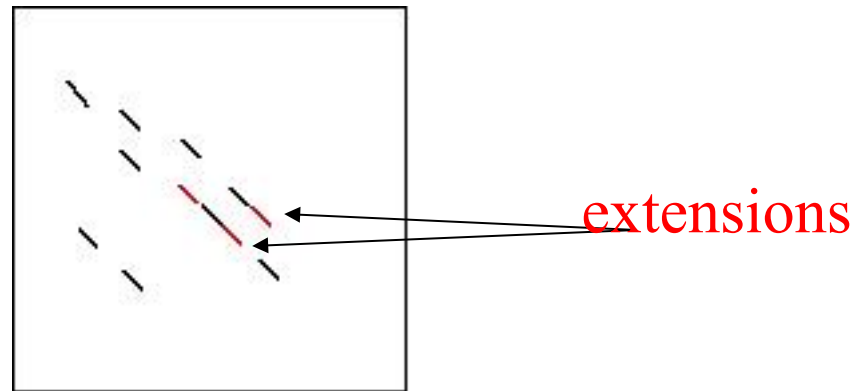


Threshold for hits, T=11

Query	1	MVRERKCILCHIVY GSK KEMDEHMRSMMLHHRELENLKGRD	40
		MVRERKCILCHI++GS+KEMDEHMRSMMLHHRELENLKGR+	
Sbjct	1	MVRERKCILCHIIH GSE KEMDEHMRSMMLHHRELENLKGRE	40

Step 3: Extending hits

- Parameter: X (controlled by a user)
- Extend the hits in both ways along diagonal (ungapped alignment) until score drops more than X relative to the best score yet attained.
- Return the score highest scoring segment pair (HSP).



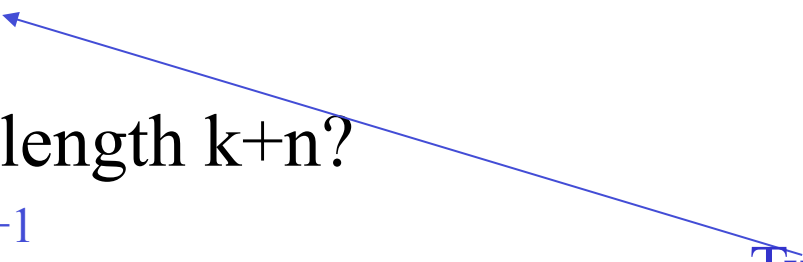
Statistical Significance of BLAST scores

Is the score high enough to provide evidence of homology?

Are the scores of alignments of random sequences higher than this score?

What are is the expected number of alignments between random sequences with score greater than this score?

BLAST statistics- intuition

- Given a 0/1 sequence of length k
 - Probability of all ones: $1/2^k$
 - Sequence of k consecutive one in a sequence length $k+1$?
 - $1 - (1-1/2^k)^2$
 - Sequence of length $k+n$?
 - $1 - (1-1/2^k)^{n+1}$
 - The longer the sequence, the more likely you are going to get k ones by chance!
- Two probes
- 

More intuition

- The probability will depend on:
 - How long is are the sequences (the longer the easier to get local score above treshold by chance)
 - Scoring matrix
 - Distribution of amino acids in each sequence

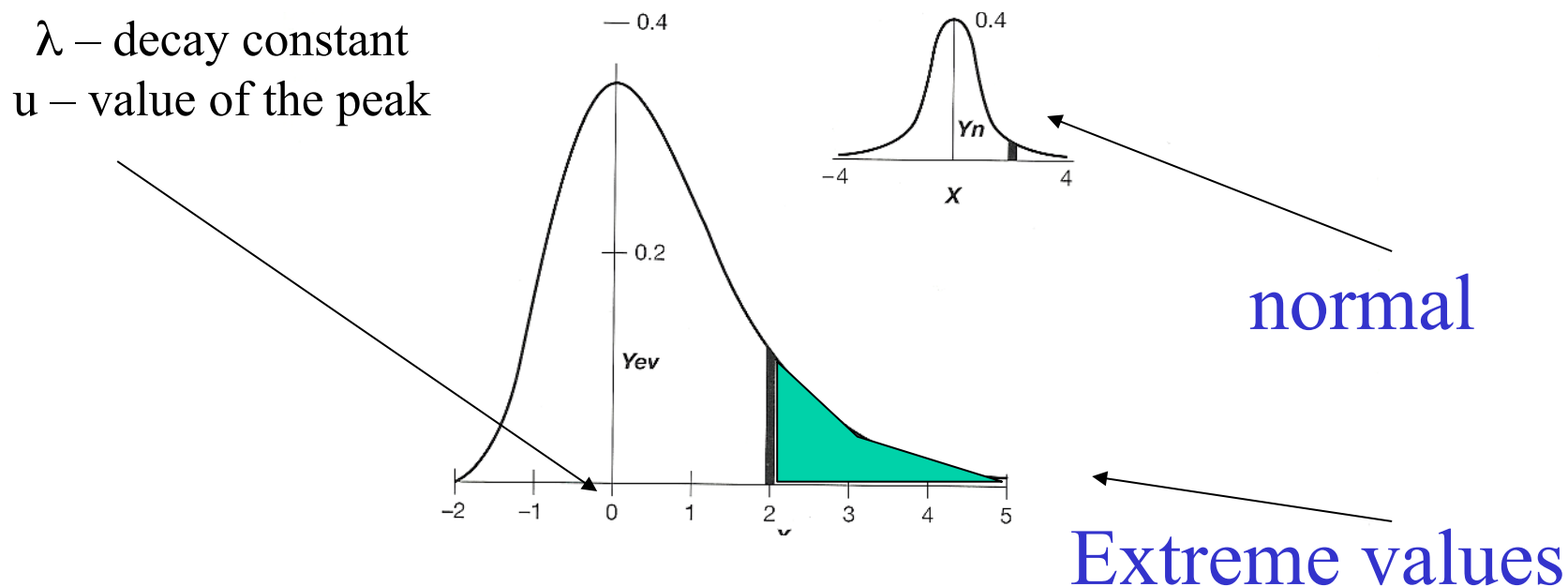
Score statistics

- If one knows the null distribution of the scores (scores of alignment of unrelated sequences) then we can assess the significance
- In order to solve this problem we will focus first on local alignments that do not contain any gaps.
- Karlin and Altschul (PNAS, 1990) provided a theory for ungapped high-scoring segments HSPs.

Karlin and Altschul provided a theory for computing such probability

- **Assumptions:**
 - the scoring matrix M must be such that the score for a random alignment is negative;
 - n, m (lengths of the aligned sequences) are large
 - The amino acid distribution $p(x)$ is in the query sequence and the data base is the same
 - Positive score is possible (i.e. M has at least one positive entry).

Score of high scoring sequence pairs follows extreme value distribution



$$P(S < x) = \exp(-e^{-\lambda(x-\mu)}) \text{ thus:}$$

$$P(S \geq x) = 1 - \exp(-e^{-\lambda(x-\mu)})$$

Extreme value distribution for sequence alignment

Property of extreme value distribution:

$$P(S < x) = \exp(-e^{-\lambda(x-\mu)}) \rightarrow$$

$$P(S \geq x) = 1 - \exp(-e^{-\lambda(x-\mu)})$$

μ – location (zero in the fig from last slide), λ scale;

For random sequence alignment:

$$\mu = \ln Kmn / \lambda$$

K - constant that depends on $p(x)$ and scoring matrix M

Since $1 - \exp(-x) \sim x$ and substituting for μ and σ :

$$P(S \geq x) \sim e^{-\lambda(x-\mu)} = Kmn e^{-\lambda x}$$

E=value-expected number of random scores above X

- E-value = $KNme^{-\lambda x}$

(Expected number of sequences scoring at least x
observed by chance, it is approximately same as
p value for p value < 0.1)

Normalization

After normalization to by setting

$$S' = (\lambda S - \ln K) / \ln 2$$

we get “bit score” S' such that

$$E = Nm 2^{-S'} \text{ (blast e-value)}$$

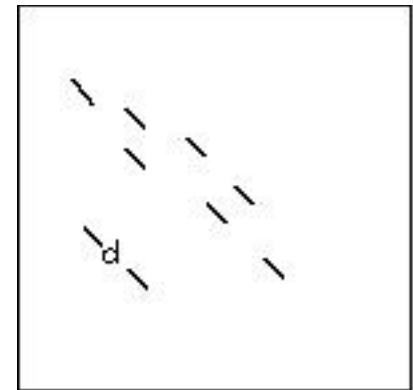
**Bit scores from various scoring matrices
can be compared directly**

For BLAST tutorial visit

<http://www.ncbi.nlm.nih.gov/BLAST/>

Refinement of the basic algorithm-the two hit method

- **Observation:** HSP of interest are long and can contain multiple hits relatively short distance away.
- **Central idea:** Look for non-overlapping pairs of hits that are of distance at most d on the same diagonal.
- **Benefits:**
 - can reduce word size w from 3 to 2 without losing sensitivity (actually sensitivity of two-hit BLAST is higher).
 - Since extending a hit requires a diagonal partner, smaller number of hits are being extended results in increased speed.



Gapped BLAST statistics

- Theory for ungapped BLAST does not extend easily
- Simulations indicate that for the best hits scores for local alignment follow extreme value distribution
- Method approximate λ and μ to match experimental distribution - λ and μ can be computed from median and variation of the experimental distribution.
- BLAST approach – simulate the distribution for set of scoring matrices and a number of gap penalties. BLAST offers choice of parameters from this pre-computed set..

NCBI HomePage - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop Search Print

Home Bookmarks mozilla.org Latest Builds Netflix - Rent DVDs ... Google NLM PubMed http://www.ncbi.nlm...

Top Up First Previous Next Last Document More



National Center for Biotechnology Information

National Library of Medicine

National Institutes of Health

PubMed All Databases BLAST OMIM Books TaxBrowser Structure

Search for

SITE MAP

Alphabetical List
Resource Guide

About NCBI

An introduction to
NCBI

GenBank

Sequence
submission support
and software

Literature databases

PubMed, OMIM,
Books, and PubMed
Central

Molecular databases

Sequences,
structures, and
taxonomy

Genomic biology

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More...](#)

100 Gigabases

GenBank and its collaborating databases, the European Molecular Biology Laboratory and the DHA Databank of Japan, have reached a milestone of 100 billion bases from over 165,000 organisms. See the [press release](#) or find more information on [GenBank](#).



The new My NCBI has replaced the Cubby and includes automatic e-mailing of search updates and filtering search results. A tab format is used for features such as limits and displaying filtered search results.

Hot Spots

- ▶ Assembly Archive
- ▶ Clusters of orthologous groups
- ▶ Coffee Break, Genes & Disease, NCBI Handbook
- ▶ Electronic PCR
- ▶ Entrez Home
- ▶ Entrez Tools
- ▶ Gene expression omnibus (GEO)
- ▶ Human genome resources
- ▶ Malaria genetics & genomics

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop <http://www.ncbi.nlm.nih.gov/BLAST/> Search Print

Home Bookmarks mozilla.org Latest Builds Netflix - Rent DVDs ... Google NLM PubMed <http://www.ncbi.nlm.nih.gov/BLAST/>

Top Up First Previous Next Last Document More

CBI → BLAST Latest news: 28 August 2005 : BLAST 2.2.12 released

Getting started

News

FAQs

info

NAR 2004

NCBI Handbook

The Statistics of Sequence Similarity Scores

are

Downloads

Developer info

resources

References

NCBI Contributors

Mailing list

Contact us

The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Nucleotide <ul style="list-style-type: none">Quickly search for highly similar sequences (megablast)Quickly search for divergent sequences (discontiguous megablast)Nucleotide-nucleotide BLAST (blastn)Search for short, nearly exact matchesSearch trace archives with megablast or discontiguous megablast	Protein <ul style="list-style-type: none">Protein-protein BLAST (blastp)Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)Search for short, nearly exact matchesSearch the conserved domain database (rpsblast)Protein homology by domain architecture (cdart)
Translated <ul style="list-style-type: none">Translated query vs. protein database (blastx)Protein query vs. translated database (tblastn)Translated query vs. translated database (tblastx)	Genomes <ul style="list-style-type: none">Human, mouse, rat, chimp NEW, cow, pig, dog, sheep, catChicken, puffer fish, zebrafishEnvironmental samplesMalariaInsects, nematodes, plants, fungi, microbial genomes, other eukaryotic genomes
Special <ul style="list-style-type: none">Search for gene expression data (GEO BLAST)	Meta <ul style="list-style-type: none">Retrieve results

NCBI
Nucleotide Protein Translations Retrieve results for an RID
protein-protein **BLAST**

[Search](#)

```
>gi|46849712|ref|NP_035873.1| SH3-domain binding protein 3  
[Mus musculus]  
MVRERKCILCHIVYGSKKEMDEHMRSMLHHRELENLKGRDISHECRVCRVTEVGLSAYAKI  
VDAQEREDDGKEEEEEYYFDKELVQLIQERKEQSRQDEPPSNSQEVNSDDRQPQWRREDR  
QPPRHHRGPPQRDVKWEKDGFNSTRKNSFPHSLRNSGGPRGSSVWHKGATRGSSSTWFLNHS
```

[Set subsequence](#) From: To:

[Choose database](#)

nr
nr
refseq
swissprot
pat
pdb
env_nr
month

[Do CD-Search](#)

Now:

Options for advanced blasting

[Limit by entrez
query](#)

or select from:

Options for advanced blasting

[Limit by entrez query](#) or select from:

[Compositional adjustments](#)

[Choose filter](#) ☒ Low complexity ☐ Mask for lookup

[Expect](#)

[Word Size](#)

[Matrix](#) Gap Costs

[PSSM](#)

[Other advanced](#)

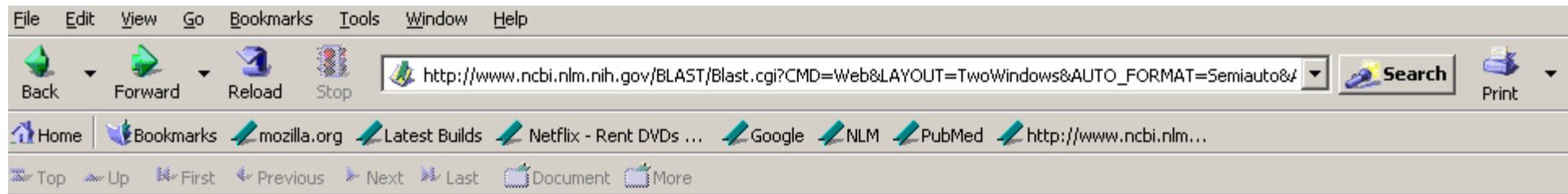
[PHI pattern](#)

All organisms [ORGN]
Viruses [ORGN]
Archaea [ORGN]
Bacteria [ORGN]
Eukaryota [ORGN]
Viridiplantae [ORGN]
Fungi [ORGN]
Metazoa [ORGN]
Arthropoda [ORGN]
Vertebrata [ORGN]
Mammalia [ORGN]
Rodentia [ORGN]
Primates [ORGN]

Existence: 11 Extension: 1
Existence: 9 Extension: 2
Existence: 8 Extension: 2
Existence: 7 Extension: 2
Existence: 12 Extension: 1
Existence: 11 Extension: 1
Existence: 10 Extension: 1

Low complexity regions

- In some protein sequences there are regions with low information content (the “low complexity regions”) – e.g. regions that contains that have a large number of, say, leucine; or repeats
- But, since BLAST assumes uniformly-distributed amino-acid sequences
- BLAST provides possibility to mask such regions:
(BLAST has the filter turn ON by default.)



Format

Show ☒ [Graphical Overview](#) ☒ [Linkout](#) ☒ [Sequence Retrieval](#) ☒ [NCBI-gi](#) Alignment in HTML format

[Masking Character](#) Default(X for protein, n for nucleotide) [Masking Color](#) Black

Number of: [Descriptions](#) 500 [Alignments](#) 250

[Alignment view](#) Pairwise

[Format for PSI-BLAST](#) ☐ [with inclusion threshold](#): 0.005

[Limit results by entrez query](#) or select from: All organisms

[Expect value range](#):

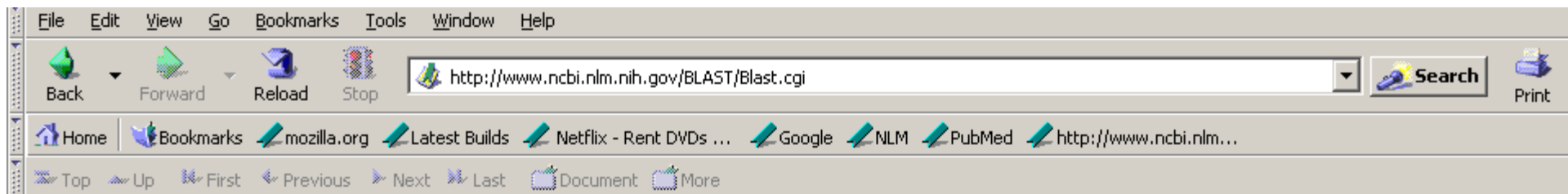
[Layout](#): Two Windows [Formatting options on page with results](#): None

[Autoformat](#) Semi-auto



Get the URL with preset values ? [Get URL](#)





Your request has been successfully submitted and put into the Blast Queue.

Query = gi|46849712|ref|NP_035873.1| SH3-domain binding protein 3 [Mus musculus] (1888 letters)

Putative conserved domains have been detected, click on the image below for detailed results.



The request ID is

Format! or **Reset all**

The results are estimated to be ready in 15 seconds but may be done sooner.

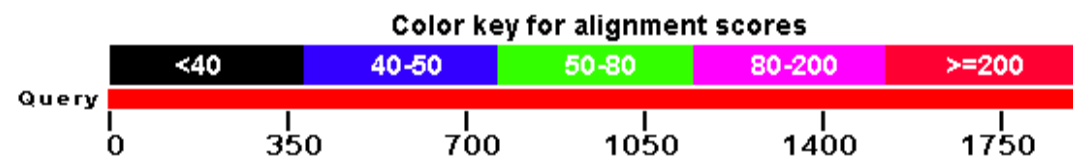
Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may request results of a different search by entering any other valid request ID to see other recent jobs.

Format



Distribution of 1586 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments



Unaligned region

Masked Regions

Descending
Score
order

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop <http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi#46849712> Search Print

Home Bookmarks mozilla.org Latest Builds Netflix - Rent DVDs ... Google NLM PubMed <http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi#46849712>

Top Up First Previous Next Last Document More

> [gi|46849712|ref|NP_035873.1](#) **G** SH3-domain binding protein 3 [Mus musculus]
[gi|3372657|gb|AAD04329.1](#) **G** zinc finger protein 106 [Mus musculus]
Length=1888

Score = 3506 bits (9091), Expect = 0.0
Identities = 1888/1888 (100%), Positives = 1888/1888 (100%), Gaps = 0/1888 (0%)

Query	1	MVRERKCILCHIVYGSKKEMDEHMRSMLHHRELENLKGRDISHECRVCRVTEVGLSAYAK	60
		MVRERKCILCHIVYGSKKEMDEHMRSMLHHRELENLKGRDISHECRVCRVTEVGLSAYAK	
Sbjct	1	MVRERKCILCHIVYGSKKEMDEHMRSMLHHRELENLKGRDISHECRVCRVTEVGLSAYAK	60
Query	61	HISGQLHKDNVDAQXXXXXXXXXXXXXXXXXXLVQLIQERKEQSRQDEPPSNSQEVNSDD	120
		HISGQLHKDNVDAQEREDDGKEEEEEEEYFDKELVQLIQERKEQSRQDEPPSNSQEVNSDD	
Sbjct	61	HISGQLHKDNVDAQEREDDGKEEEEEEEYFDKELVQLIQERKEQSRQDEPPSNSQEVNSDD	120
Query	121	RQPQWRREDRIPYQDRESYSXXXXXXXXXXXXXDWKWEKDGFNSTRKNSFPHSLRNSGGPR	180
		RQPQWRREDRIPYQDRESYSQPPRHHRGPPQRDWKWEKDGFNSTRKNSFPHSLRNSGGPR	
Sbjct	121	RQPQWRREDRIPYQDRESYSQPPRHHRGPPQRDWKWEKDGFNSTRKNSFPHSLRNSGGPR	180
Query	181	GSSVWHKGATRGSSSTWFLXXXXXXXXXXXXXXXXXWVDWNYNGTGRNSSWHSEGTTGGFPWUHM	240
		GSSVWHKGATRGSSSTWFLNHSNSGGGWHSNNGMVDWNYNGTGRNSSWHSEGTTGGFPWUHM	
Sbjct	181	GSSVWHKGATRGSSSTWFLNHSNSGGGWHSNNGMVDWNYNGTGRNSSWHSEGTTGGFPWUHM	240
Query	241	NNSNGNWKSSVRSTNSWNYNGPGDKFQQGRNRNPNYQMEDMTKMWNKKSNKPSKYSQERC	300
		NNSNGNWKSSVRSTNSWNYNGPGDKFQQGRNRNPNYQMEDMTKMWNKKSNKPSKYSQERC	
Sbjct	241	NNSNGNWKSSVRSTNSWNYNGPGDKFQQGRNRNPNYQMEDMTKMWNKKSNKPSKYSQERC	300
Query	301	KWQRQDRDKAAKYRSPPEGYASDTFPSEGLLEFNFEQRESQTTKQTDTAASKINGKNGTK	360
		KWQRQDRDKAAKYRSPPEGYASDTFPSEGLLEFNFEQRESQTTKQTDTAASKINGKNGTK	
Sbjct	301	KWQRQDRDKAAKYRSPPEGYASDTFPSEGLLEFNFEQRESQTTKQTDTAASKINGKNGTK	360
Query	361	ARDKFRRWTPYPSQKTLDLQSALKEVIGSKSDTLEKPLFNFLITAGLRKPVDKTSNPPV	420
		ARDKFRRWTPYPSQKTLDLQSALKEVIGSKSDTLEKPLFNFLITAGLRKPVDKTSNPPV	
Sbjct	361	ARDKFRRWTPYPSQKTLDLQSALKEVIGSKSDTLEKPLFNFLITAGLRKPVDKTSNPPV	420
Query	421	LYTQYACDPCGDCWIAICDCTACGELDADCTTCGCEHICGCMFLXXXXXXXXXXXXXXXXXX	480

Done

Some rules of thumb

- Significant hits for **protein** searches:
E-value $\leq 1e^{-03}$
Percent of identity $\geq 25\%$
- Significant hits for **nucleotide** searches:
E-value $\leq 10^{-06}$
Percent of identity $\geq 70\%$

Variants BLAST Algorithms:

Program	Query	Database
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nucleotide, six-frame translation	Protein
TBLASTN	Protein	Nucleotide, six-frame translation
TBLASTX	Nucleotide, six-frame translation	Nucleotide, six-frame translation

Position Specific Iterated BLAST

- Collect all database sequence segments that have been aligned with query sequence with E-value below set threshold (default 0.01)
- 1. Construct position specific scoring matrix for collected sequences. Rough idea:
 - Align all sequences to the query sequence as the template.
 - Assign weights to the sequences
 - Construct position specific scoring matrix
- 2. Find sequences that mach the profile
- Iterate (1) and (2)

Sequence to run an example

```
LSADQISTVQASF DKVKGDPVGILYAVFKA31DPSIMAKFTQFAGKDL ESIKGT A  
PFETHAN61RIVGFFSKIIGELPNIEADVNTFVASHKPR91GVTHDQLNNFRA  
GFVSYMKAHTDFAGAEAA121WGATLDTFFGMIFSKM
```

FASTA

Heuristic algorithm, similar to BLAST.

Main idea (expanded on next slides):

- **Step 1** : Find hot-spots (hot spot ~ hit in BLAST)
- **Step 2**: Locate best “diagonal runs” (sequences of consecutive hot spots on a diagonal)
- **Step 3** : Combine sub-alignments form diagonal runs into a longer alignment
- **Step 4**: Find alternative local alignments.

Step 1 FASTA

- A Lookop table is used to find identities (ktup=1) or runs of identities

Lookup table
for sequence 1:

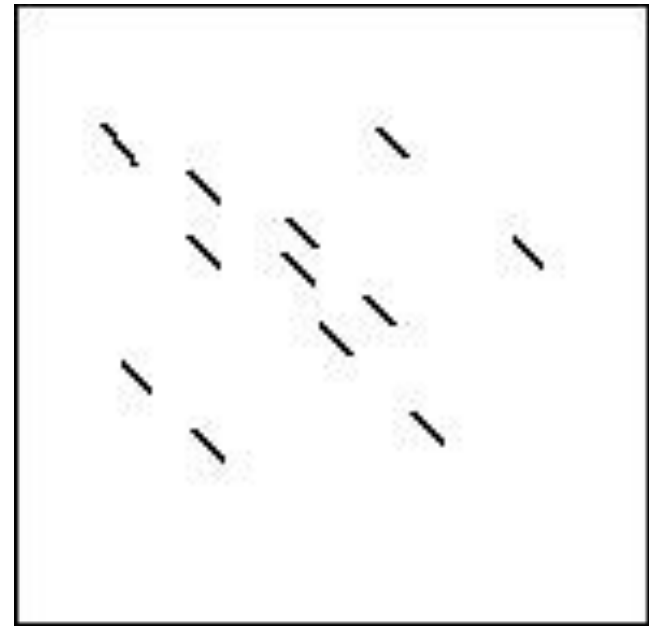
Step 2 of FASTA

Locate best diagonal runs (gapless alignments) Give positive score for each hot spot

- Give negative score for each space between hot spots
- Find best scoring runs
- Score the alignments from the runs and find ones above a threshold. These are possible “sub-alignments”

Step 3 of FASTA

- Combine sub-alignments into one alignment.
- We need to solve a problem known as the **chaining problem** : find a collection of non-contradicting sub-alignments that maximize some scoring function.
- Problem reduces to a problem close to maximum common subsequence.



Step 4 of FASTA

Find alternative local alignments

- Use dynamic programming restricted to a ribbon along the diagonal containing best run found in step 3.

Statistical significance estimation (in the absence rigorous theoretical model)

- Collect alignment scores of this sequence to other random sequence (exclude extremes)
- Compute average score, (ave.) and standard deviation, (sdiv).
- Compute z-score:

$$Z = (\text{score} - \text{ave score}) / \text{sdiv}$$

- Estimate $P(Z > z)$ (under the assumption of extreme value distribution)

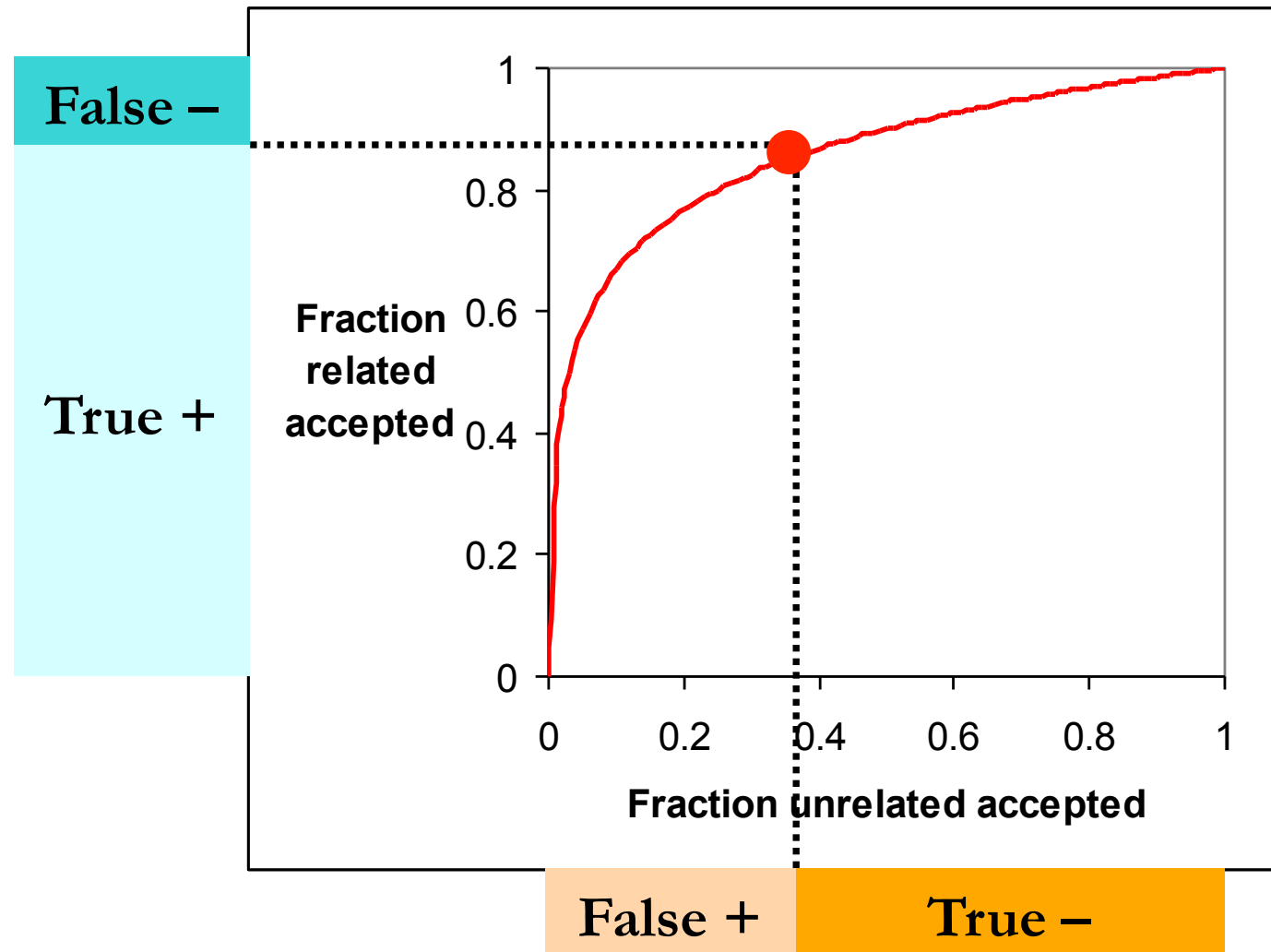
Comparing methods' retrieval accuracy

- Let's assume we have a new method to perform a search and we would like to compare with BLAST and FASTA, or just BLAST Vs. FASTA.
- First, we need to create a gold standard (of correct answers) for benchmarking (for example proteins known to be homologous based on structure comparison.)
- **Idea:** For each estimate how many answer it get wrong.
- **Problem:** The answer depends in the score threshold: for example setting high score threshold we are unlikely to recover any non-homolog but we are likely to miss a lot of homolog's
- Thus we have two types of errors: false positives and false negatives and both have to be taken into account in a comparison.

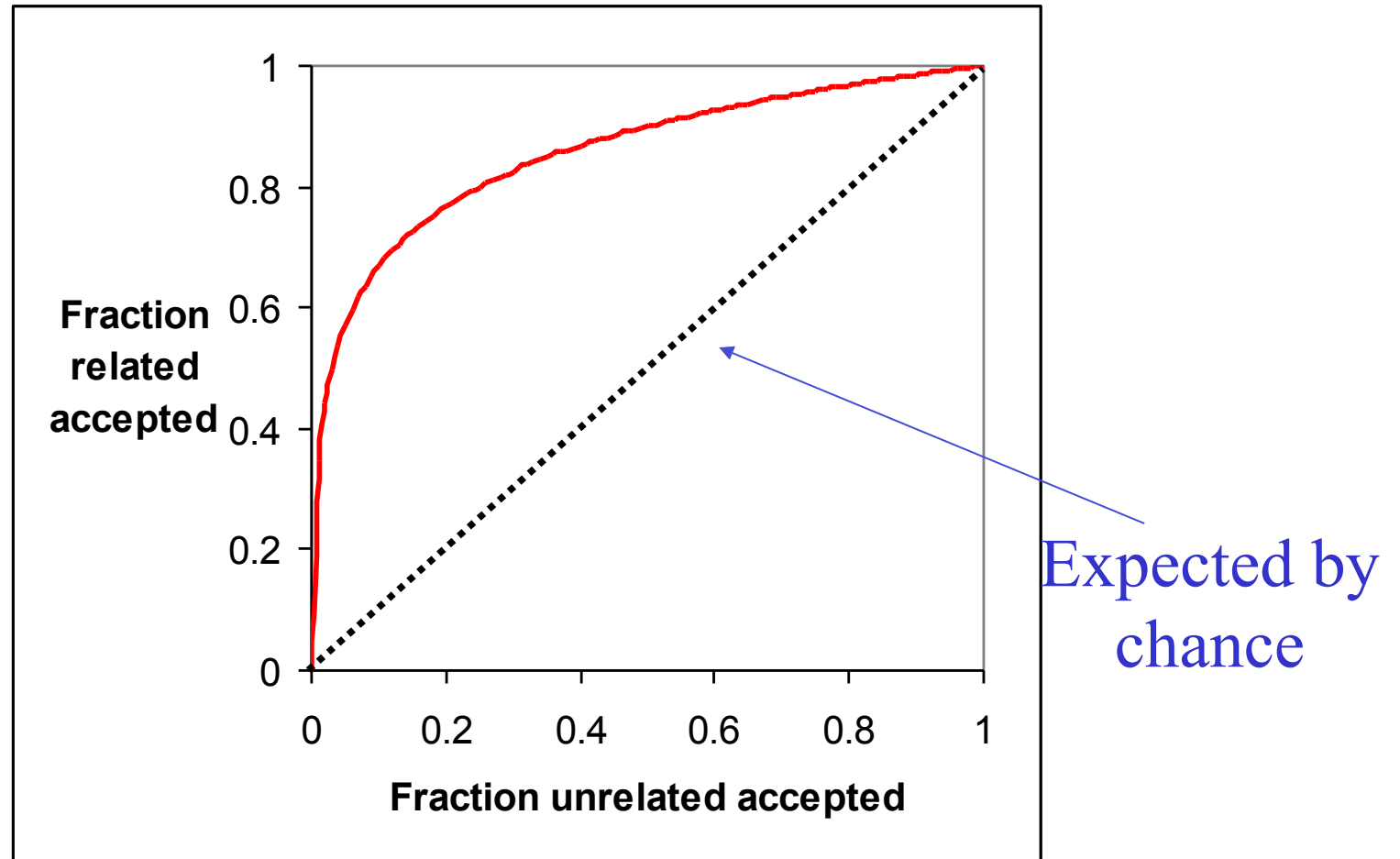
Sensitivity /Specificity of a data base search

	Related	Unrelated	Predictive value
Retrieved by the search	TP True Positive	FP False Positive	Positive (PPV) $TP/(TP+FP)$
Not retrieved by the search	FN False Negative	TN True Negative	Negative (NPV) $TN/(TN+FN)$
	Sensitivity $TP/(TP+FN)$	Specificity $TN/(FP+TN)$	

Receiver Operating Characteristic curve



Random retrieval on a *ROC* plot



ROC curve

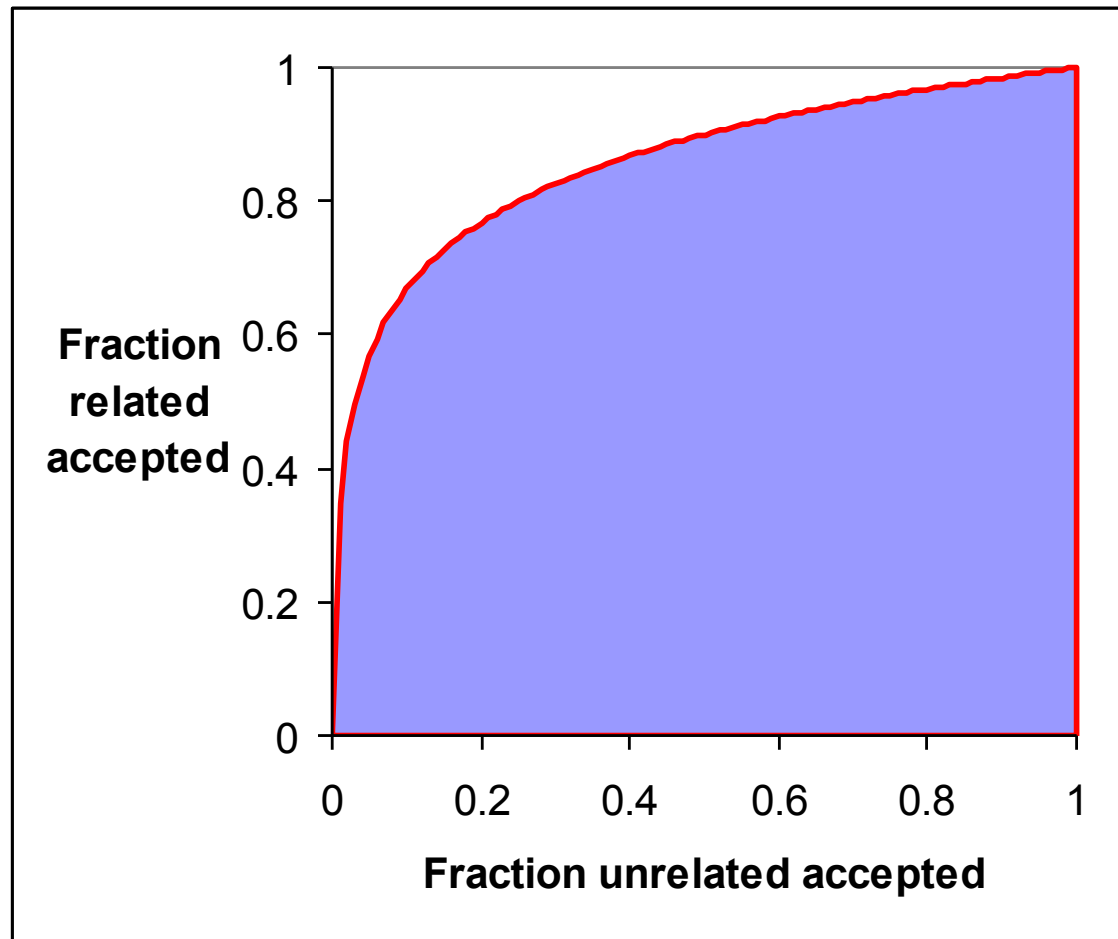
- Axis correlate with statistical measures:
- Sensitivity of the search= $TP/(TP+FN)$
- Specificity of the search= $TN/(FP+TN)$
- So ROC plots are plots of
- Sensitivity Vs. (1-Specificity)

Comment:

Other measurements are used to do other variants of this plot

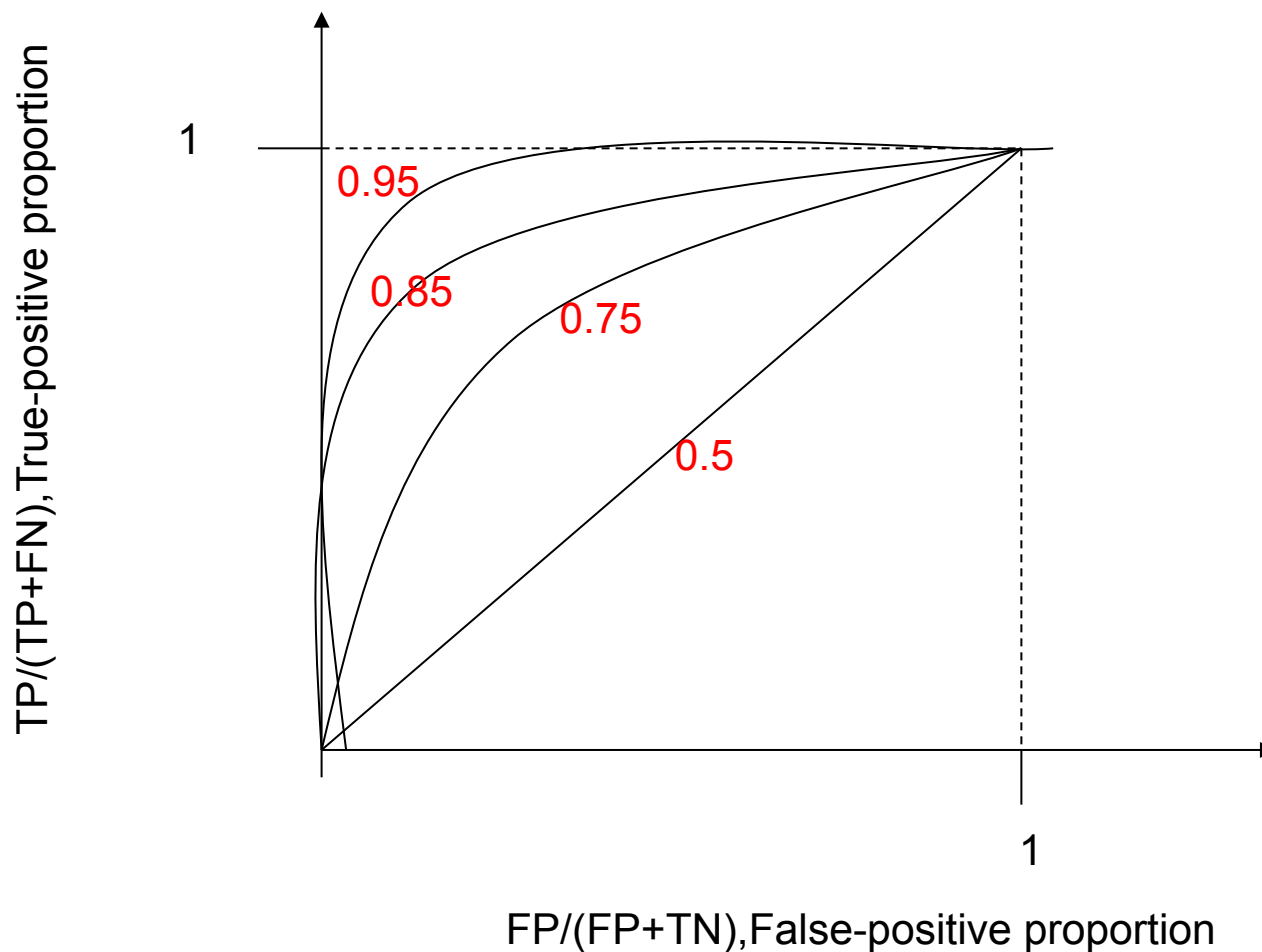
This slide is by Stephen Altschul from talk: www.dimacs.rutgers.edu/workshops/proteindomains/dimacstalk1.ppt

ROC score: area under the *ROC* curve



ROC scores – examples

better method – higher ROC score



ROC_{*n*}

If the data base is huge but the set of true positives is small you one is often interested in how many true positives are recovered before you get a certain number of false positives.

That is you are not necessarily interested in what is the order of true and false positives after a certain number of errors (*n*)

ROC_n

Let $i = 1, 2, 3 \dots$ index the rank of the false positives, and let t_i be the number of true positives ranked ahead of the i th false positive.

ROC for n false positives, defined as:

$$ROC_n = \frac{1}{nT} \sum_{1 \leq i \leq n} t_i$$

T is total number of true positives in the database,